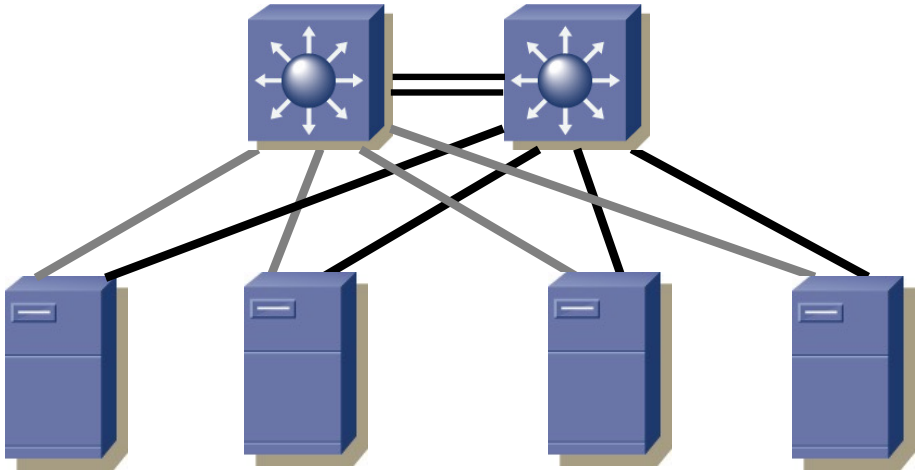
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

NICs Ethernet para servidor

Requerimientos

- Soporte de algún tipo de agregación de interfaces
  - Gestionable a nivel de usuario
  - O de sistema operativo
- Alto rendimiento
  - Arquitectura PC limitada para altas tasas de paquete/s
- Virtualización
  - Soporte eficiente de VMs en el host (más adelante)
- Almacenamiento
  - Integración con almacenamiento en red (más adelante)

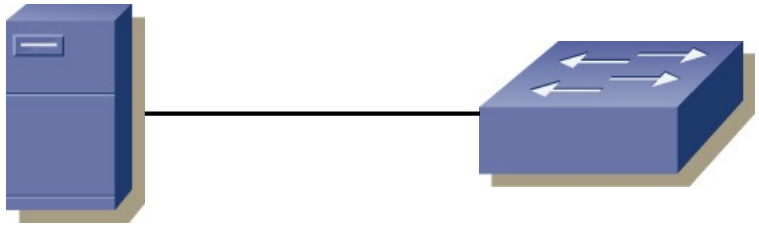


upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

Server multihoming

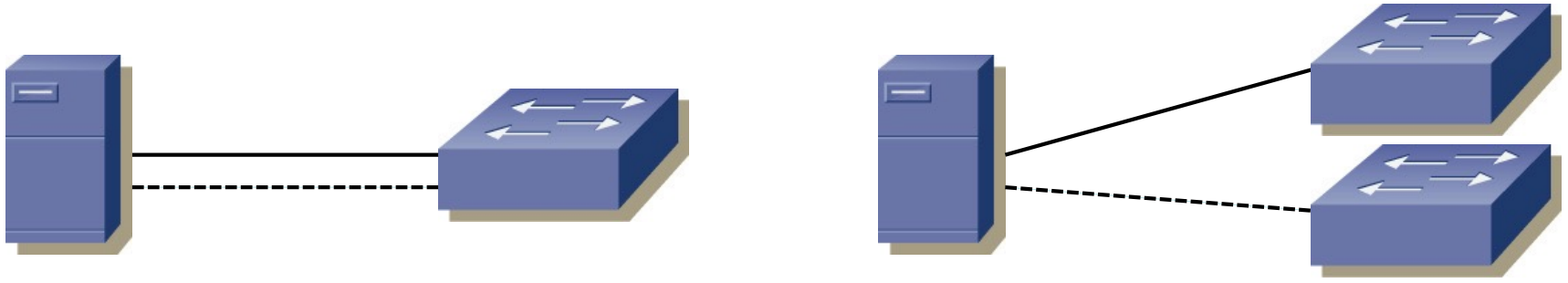
Server multihoming

- *NIC teaming / bonding / aggregation*
- Un servidor conectado a un conmutador presenta puntos únicos de fallo: la NIC, el cable, el conmutador
- Estas soluciones requieren colaboración del driver y normalmente también del sistema operativo
- Tenemos varias mejoras posibles (con una segunda o más NICs)
- (…)



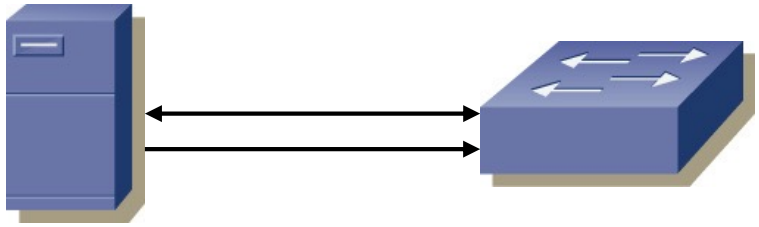
Server multihoming

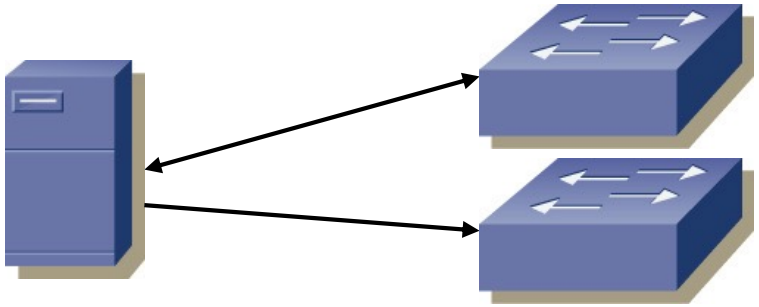
- Un segundo enlace, modo activo-pasivo
  - Si falla el primero (la NIC, el conmutador o el cable) se activa el segundo con la misma dirección MAC e IP
  - Se desaprovecha el segundo enlace



Server multihoming

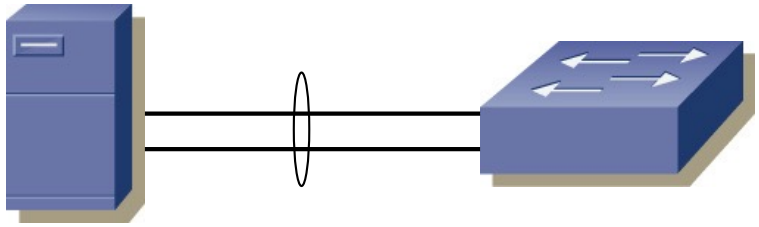
- Un segundo enlace, modo activo-pasivo
- O se usan los dos enlaces para transmitir pero solo se recibe por uno
- Cada interfaz suele enviar con diferente dirección MAC origen para no tener *MAC flapping* en el conmutador

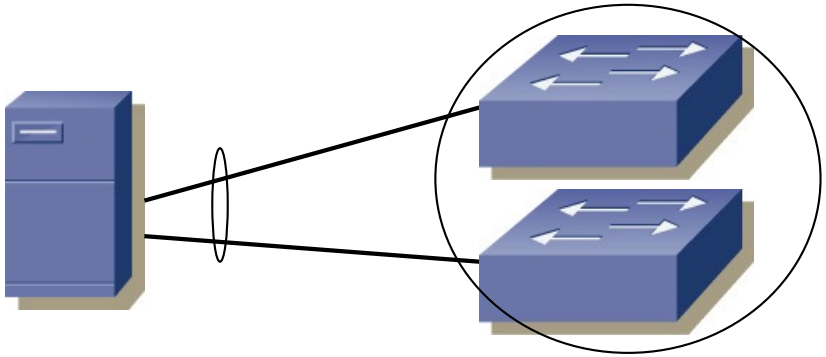




Server multihoming

- Un segundo enlace, modo activo-pasivo
- O se usan los dos enlaces para transmitir pero solo se recibe por uno
- O se forma un LAG (802.3ad / 802.1AX)
  - Permite usar la capacidad de ambos enlaces
  - Normalmente requiere colaboración por parte del switch
  - Si se quiere redundancia de switch hay que hacer una agregación en la que un extremo son 2 conmutadores

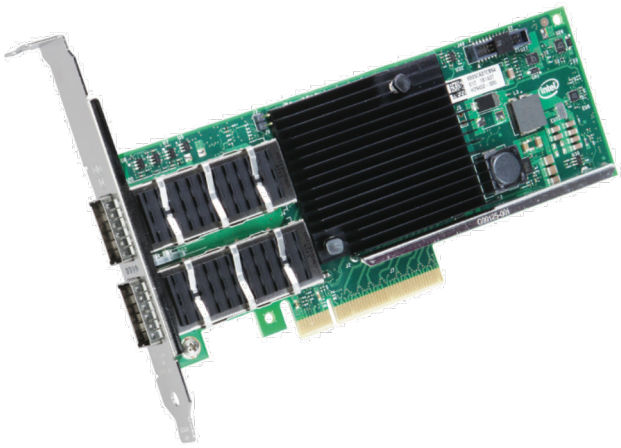




Alto rendimiento

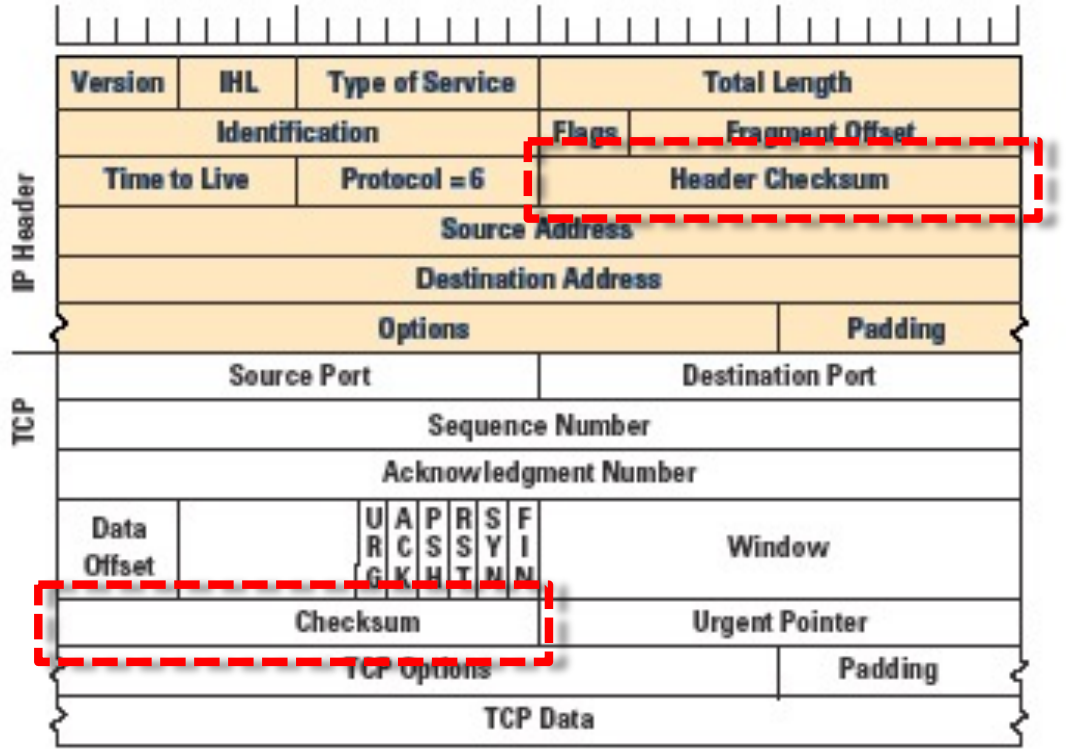
Tareas en la NIC

- Por un enlace 10GE pueden llegar en 1 segundo más de 14 millones de tramas de 64 bytes
- Eso da a la CPU unos 67ns para procesar cada una
- Las CPUs tienen serios problemas para procesar en ese tiempo cabeceras TCP/IP
- Una NIC puede incluir electrónica para llevar a cabo ciertas tareas de TCP/IP descargando a la CPU
- La NIC puede incluir ASICs, Network Processors o un procesador con un sistema operativo de tiempo real
- A 400Gbps una trama cada 1,67ns lo cual está en el rango de los mejores tiempos de acceso a memoria



Checksum offload

- La NIC descarga del cálculo a la CPU
- En transmisión y recepción
- Checksum IP (v4 y v6), UDP y TCP

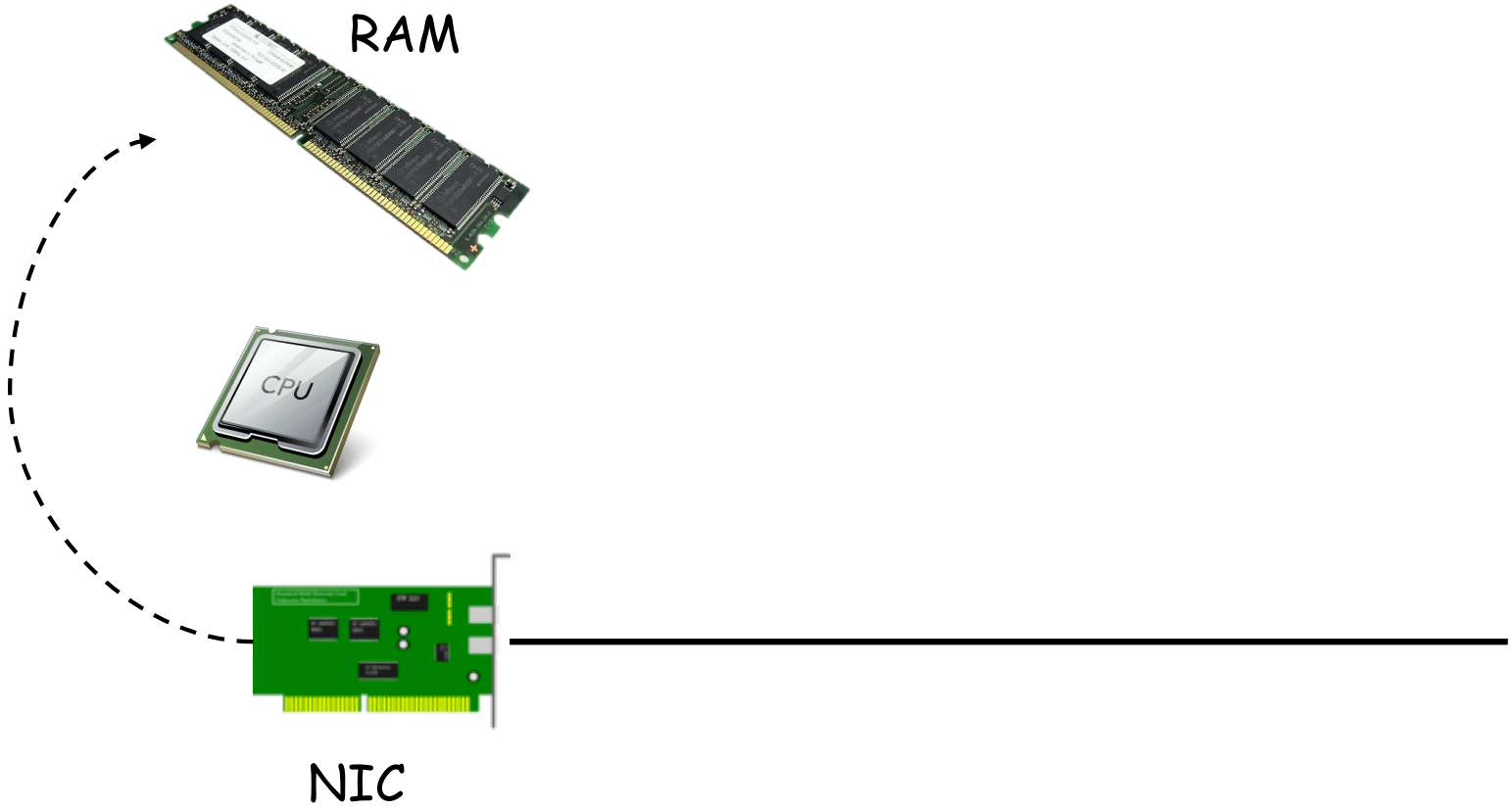


Integración en el bus

- Coalescencia de interrupciones
  - Las NICs solían generar una interrupción por paquete
  - Alto coste para la CPU
  - Por ejemplo los mainframes tienen CPUs dedicadas a atender I/O
  - La coalescencia hace que la NIC genere una interrupción para un grupo de paquetes en vez de por cada uno
  - También puede hacer *polling* la NIC

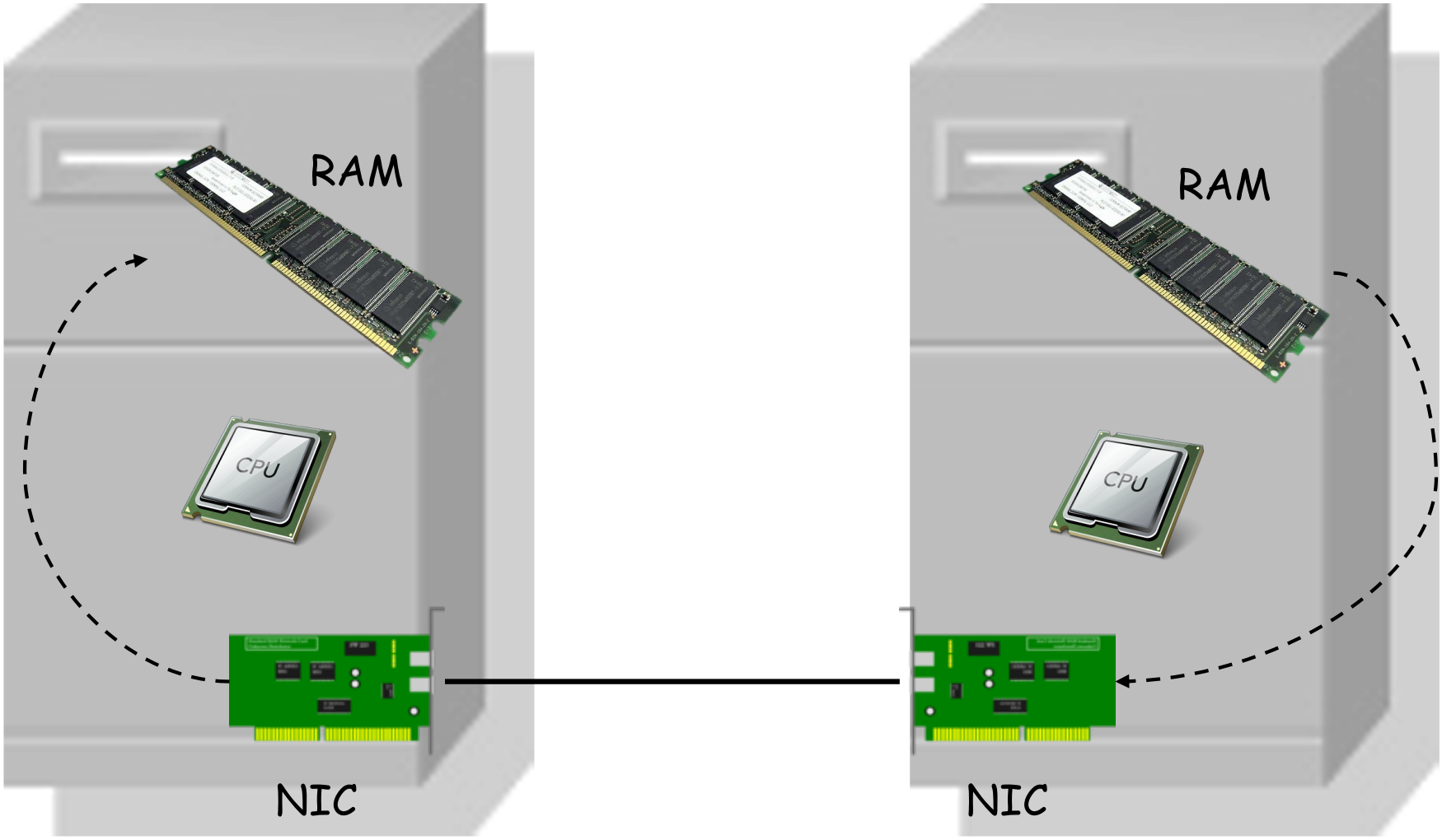
Integración en el bus

- DMA
  - *Direct Memory Access*
  - Transferencia desde la NIC a memoria sin requerir a la CPU



RDMA

- Remote Direct Memory Access
- Copias entre RAM de hosts diferentes sin involucrar a la CPU
- Latencia de pocos microsegundos



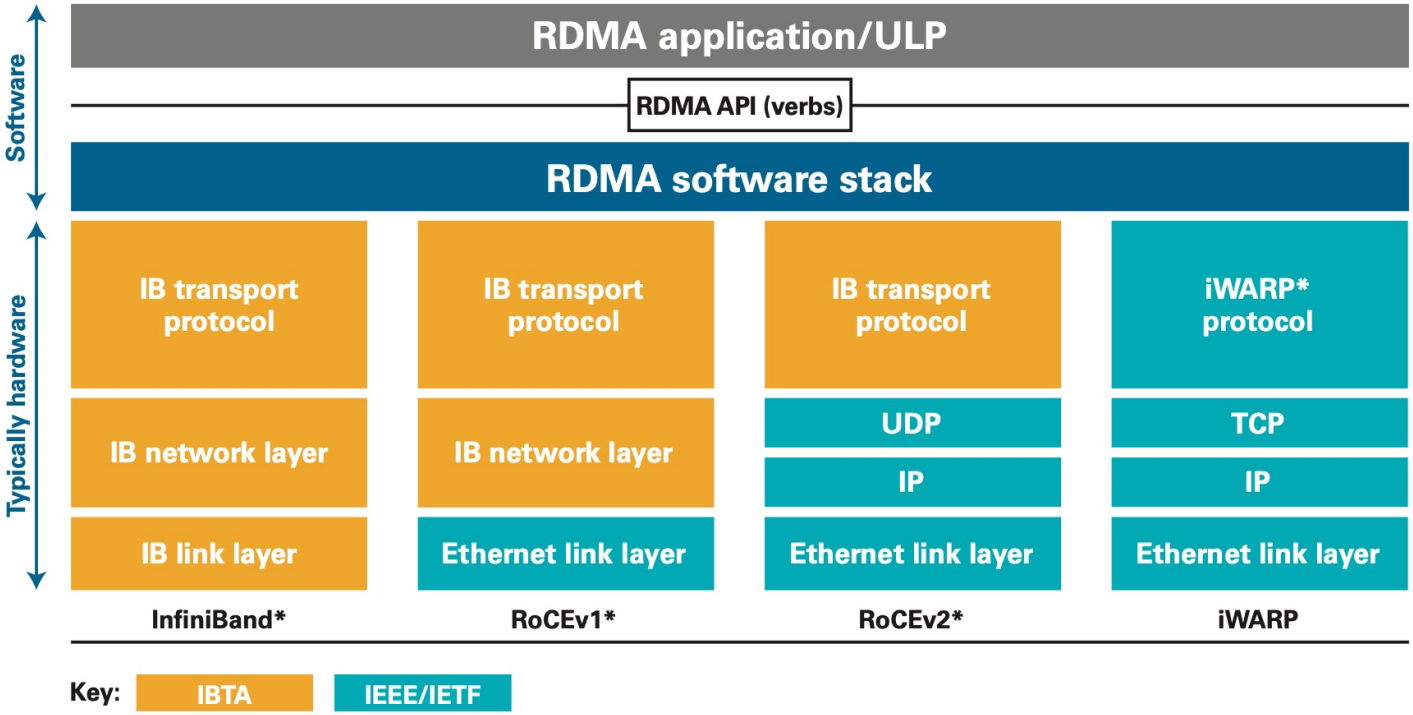
RDMA

iWARP

- RFCs 5040, 5041, 5044
- Sobre TCP o SCTP

RoCE

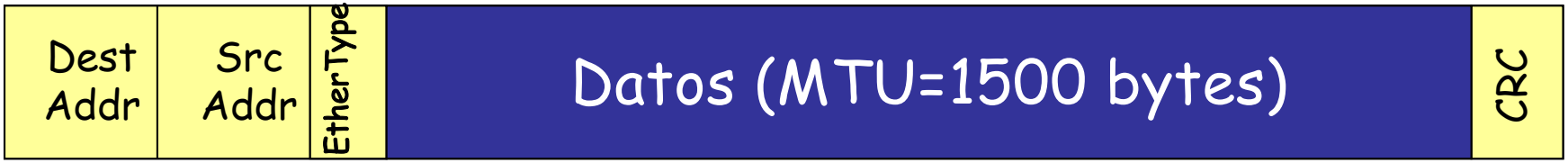
- RDMA over Converged Ethernet (DCB, Data Center Bridging)
- RoCE v1 sobre Ethernet, v2 sobre UDP
- RoCE v1 mecanismos de control de flujo y congestión de DCB
- RoCE v2 emplea control de congestión basado en ECN



https://www.intel.com/content/www/us/en/products/docs/network-io/ethernet/iwarp-rdma-here-and-now-technology-brief.html

Jumbo frames

- Tramas Ethernet con MTU superior a 1500byes
- No están estandarizadas, la MTU estándar sigue siendo de 1500bytes
- Motivos para limitarlo
  - NICs tenían memoria limitada
  - Se quería limitar el tiempo que una estación tenía capturado el medio transmitiendo
  - El CRC es menos efectivo cuanto más grande es la trama
- Hoy en día no son problemas reales:
  - Decenas o centenares de Megabytes en la NIC
  - No tenemos medio compartido (ni coaxial ni hubs)
  - El CRC de Ethernet soporta más de 11 Kbytes de trama

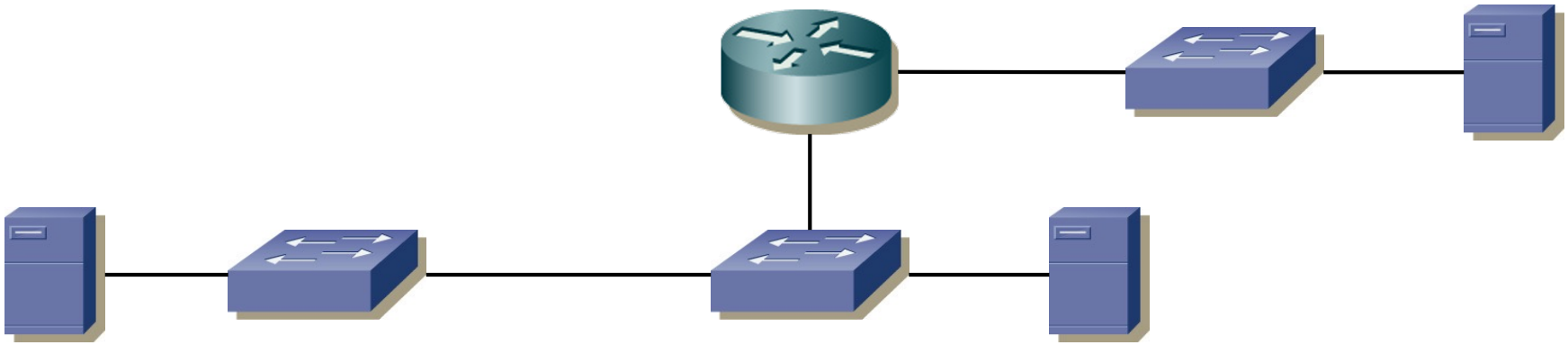


Jumbo frames

- Diversos estándares han ido aumentando el tamaño máximo de la trama (802.1Q, 802.1ad, MPLS, FCoE, etc)
- A estas últimas en ocasiones se las llama “Baby Giant”
- Jumbo frames suelen estar cerca de los 9 Kbytes (que se puedan transportar bloques de datos de 8Kbytes + encapsulados varios)
- ¿Postivo?
  - Cuanto más grandes menor ratio de cabeceras y menos interrupciones
  - Menos carga de procesado de cabeceras en equipos de red y hosts
- ¿Negativo?
  - (…)

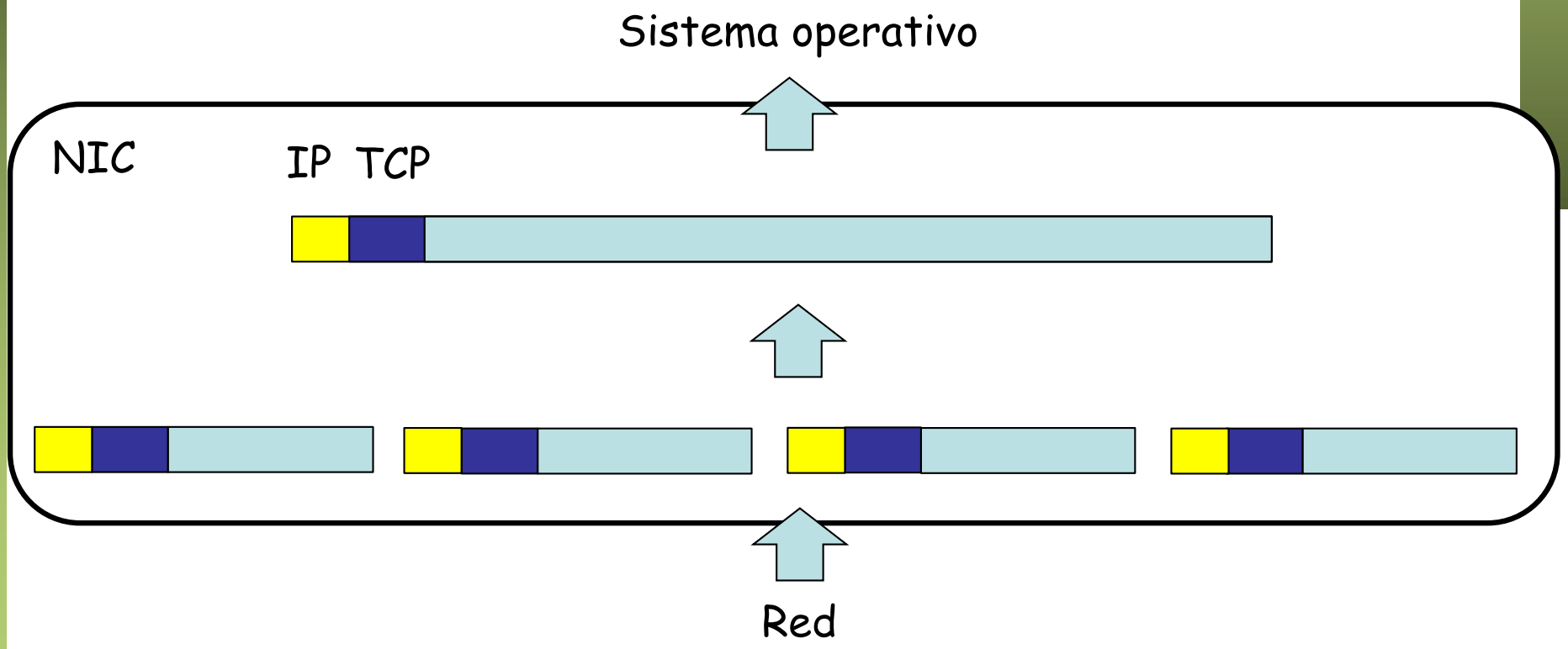
Jumbo frames

- Diversos estándares han ido aumentando el tamaño máximo de la trama (802.1Q, 802.1ad, MPLS, FCoE, etc)
- A estas últimas en ocasiones se las llama “Baby Giant”
- Jumbo frames suelen estar cerca de los 9 Kbytes (que se puedan transportar bloques de datos de 8Kbytes + encapsulados varios)
- ¿Postivo?
- ¿Negativo?
  - Todos los equipos del camino deben soportarlas
  - Posibles problemas con implementaciones que esperan 1500 bytes
  - Mayores tramas sufren mayor retardo así que no son adecuadas para todos los servicios
  - Mayores tramas pueden llenar antes los buffers de los conmutadores

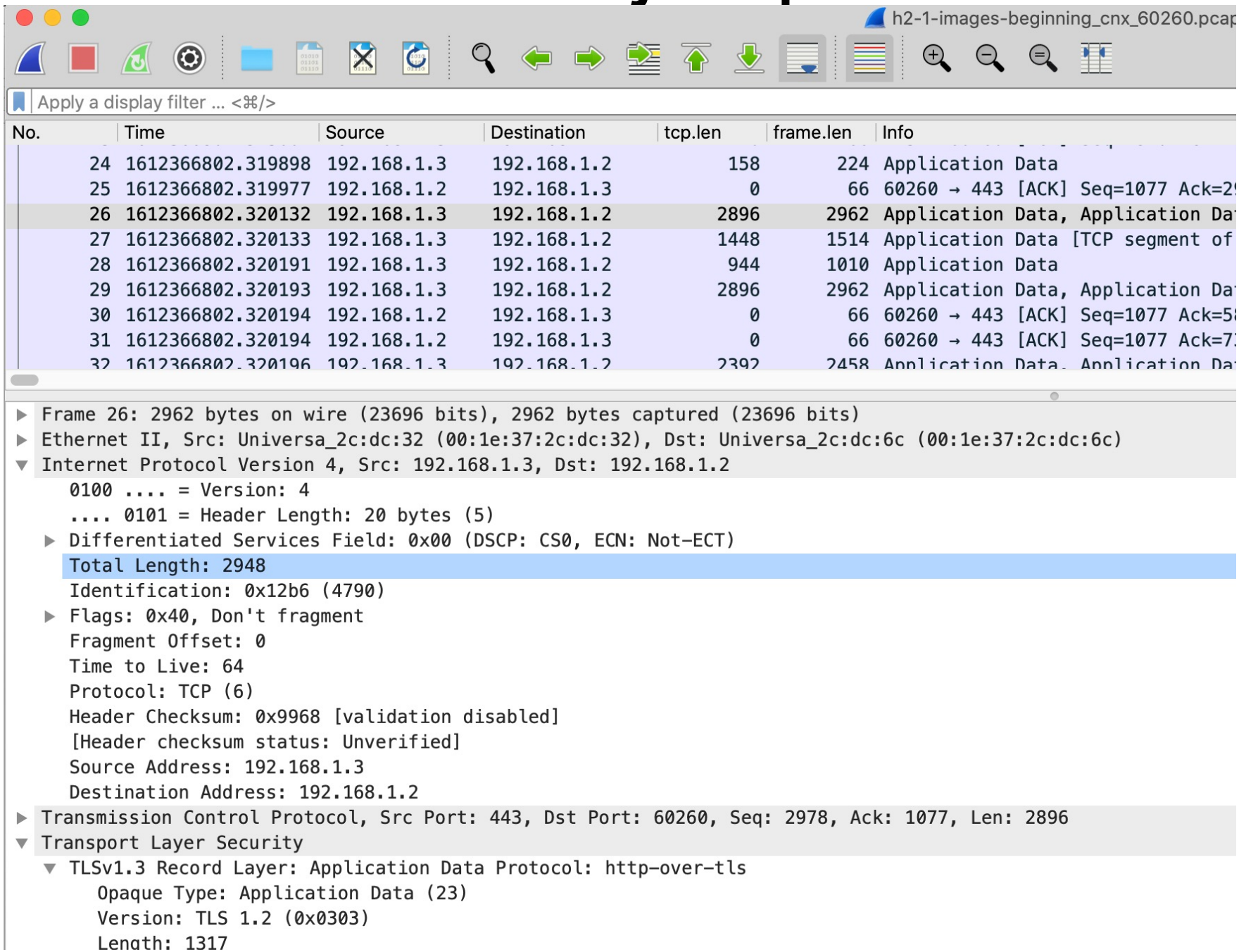


LRO

- *Large Receive Offload, Receive Segment Coalescing*
- La NIC une varios segmentos TCP en uno solo
- Crea unas cabeceras TCP e IP para ese nuevo segmento
- Reduce el número de interrupciones y procesado de cabeceras en el kernel



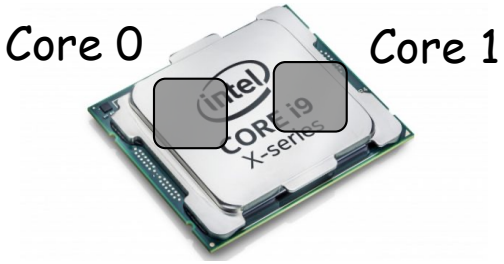
LRO: Ejemplo

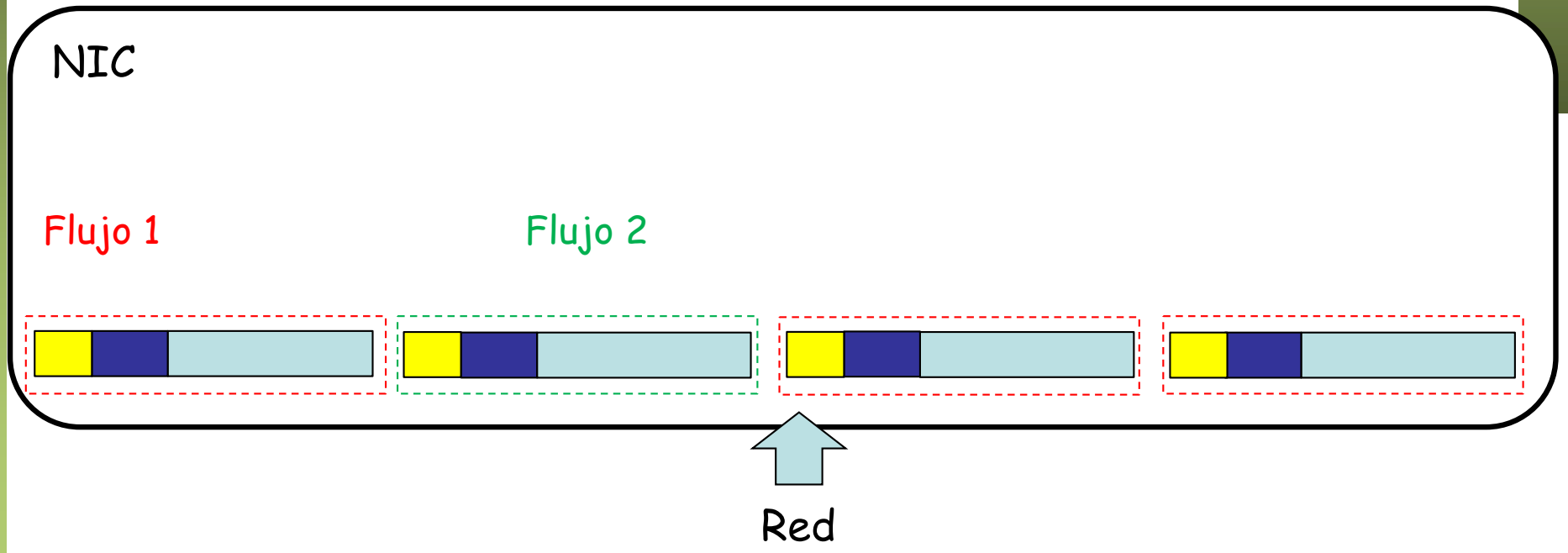

h2-1-images-beginning_cnx_60260.pcap
Apply a display filter ... <⌘/>
No. Time Source Destination tcp.len frame.len Info
24 1612366802.319898 192.168.1.3 192.168.1.2 158 224 Application Data
25 1612366802.319977 192.168.1.2 192.168.1.3 0 66 60260 → 443 [ACK] Seq=1077 Ack=2
26 1612366802.320132 192.168.1.3 192.168.1.2 2896 2962 Application Data, Application Da
27 1612366802.320133 192.168.1.3 192.168.1.2 1448 1514 Application Data [TCP segment of
28 1612366802.320191 192.168.1.3 192.168.1.2 944 1010 Application Data
29 1612366802.320193 192.168.1.3 192.168.1.2 2896 2962 Application Data, Application Da
30 1612366802.320194 192.168.1.2 192.168.1.3 0 66 60260 → 443 [ACK] Seq=1077 Ack=5
31 1612366802.320194 192.168.1.2 192.168.1.3 0 66 60260 → 443 [ACK] Seq=1077 Ack=7
Frame 26: 2962 bytes on wire (23696 bits), 2962 bytes captured (23696 bits)
Ethernet II, Src: Universa_2c:dc:32 (00:1e:37:2c:dc:32), Dst: Universa_2c:dc:6c (00:1e:37:2c:dc:6c)
Internet Protocol Version 4, Src: 192.168.1.3, Dst: 192.168.1.2
0100 .... = Version: 4
.... 0101 = Header Length: 20 bytes (5)
Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
Total Length: 2948
Identification: 0x12b6 (4790)
Flags: 0x40, Don't fragment
Fragment Offset: 0
Time to Live: 64
Protocol: TCP (6)
Header Checksum: 0x9968 [validation disabled]
[Header checksum status: Unverified]
Source Address: 192.168.1.3
Destination Address: 192.168.1.2
Transmission Control Protocol, Src Port: 443, Dst Port: 60260, Seq: 2978, Ack: 1077, Len: 2896
Transport Layer Security
TLSv1.3 Record Layer: Application Data Protocol: http-over-tls
Opaque Type: Application Data (23)
Version: TLS 1.2 (0x0303)
Length: 1317

RSS

- *Receive Side Scaling*
- Multi-CPU o CPU multi-core

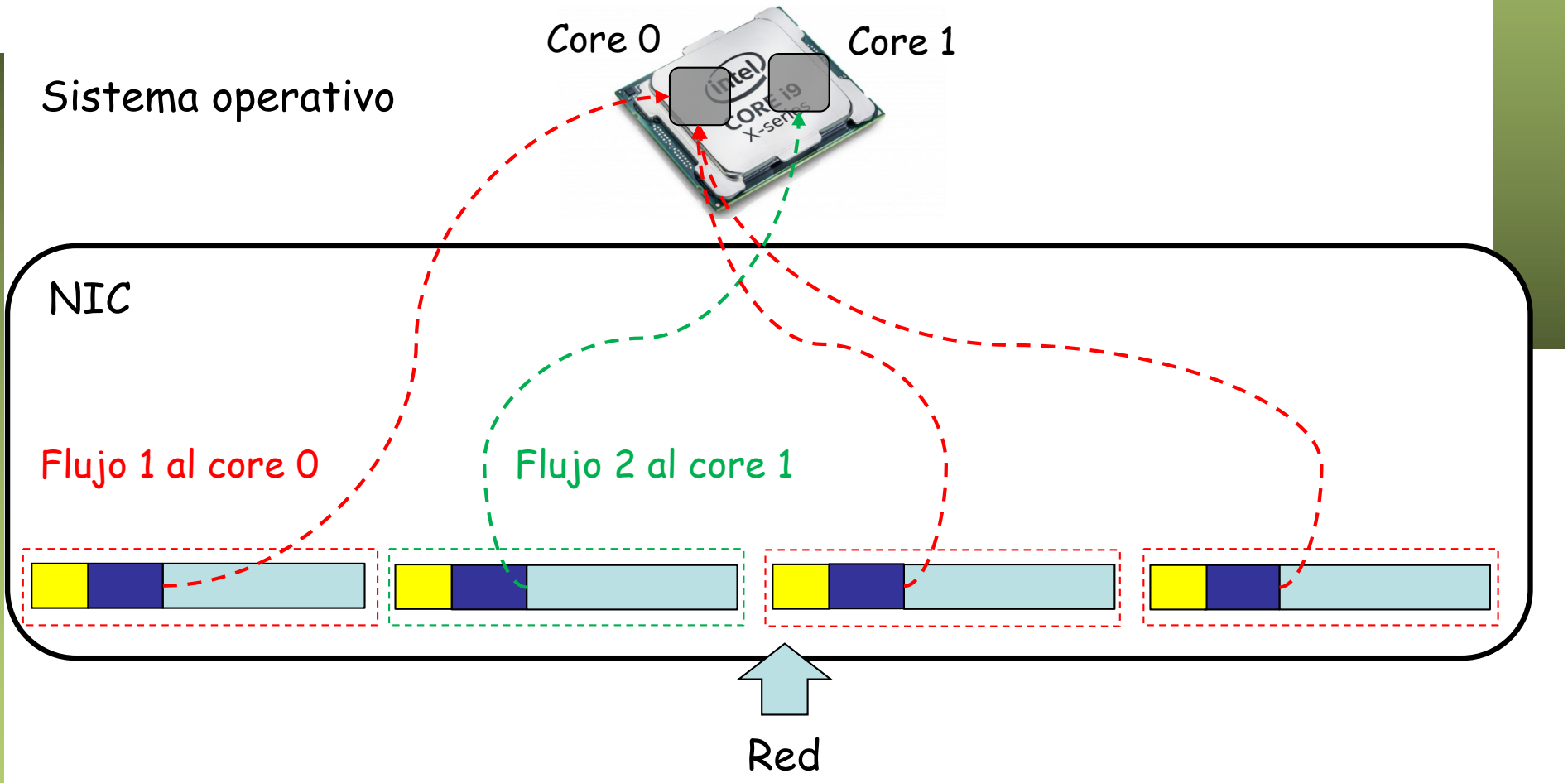
Sistema operativo





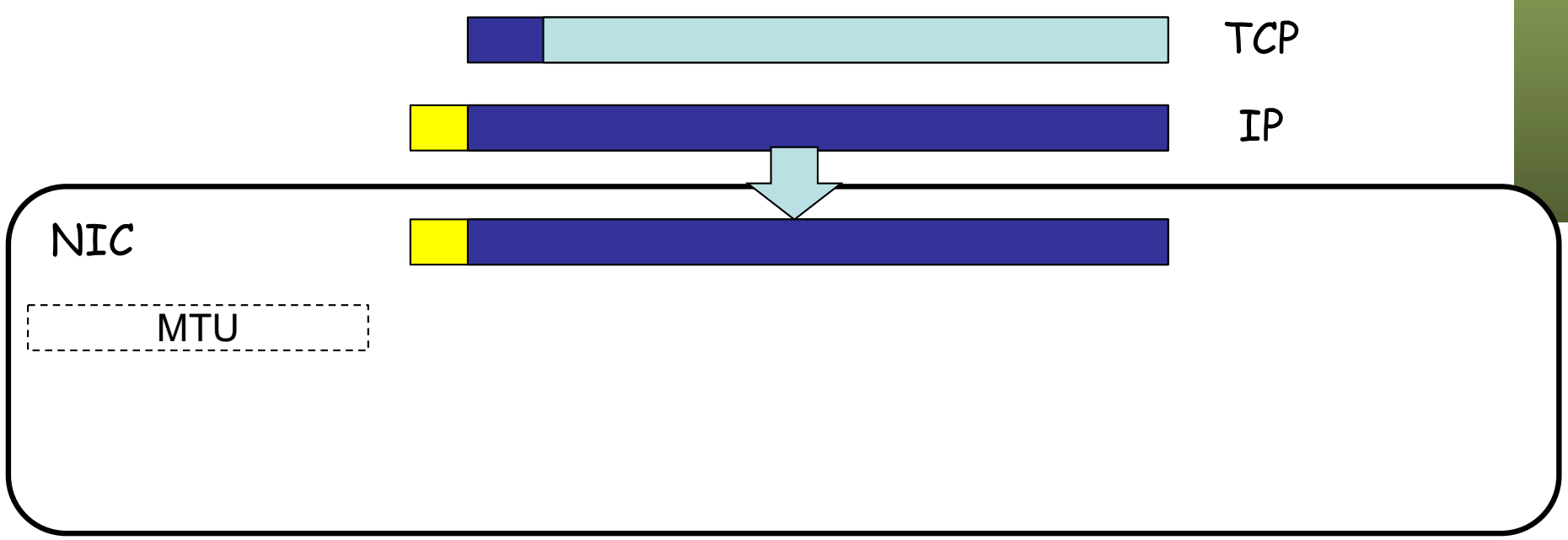
RSS

- NIC calcula un hash sobre el paquete recibido y con él decide a qué CPU manda la interrupción
- Permite paralelizar entre varias CPUs el procesado del tráfico recibido



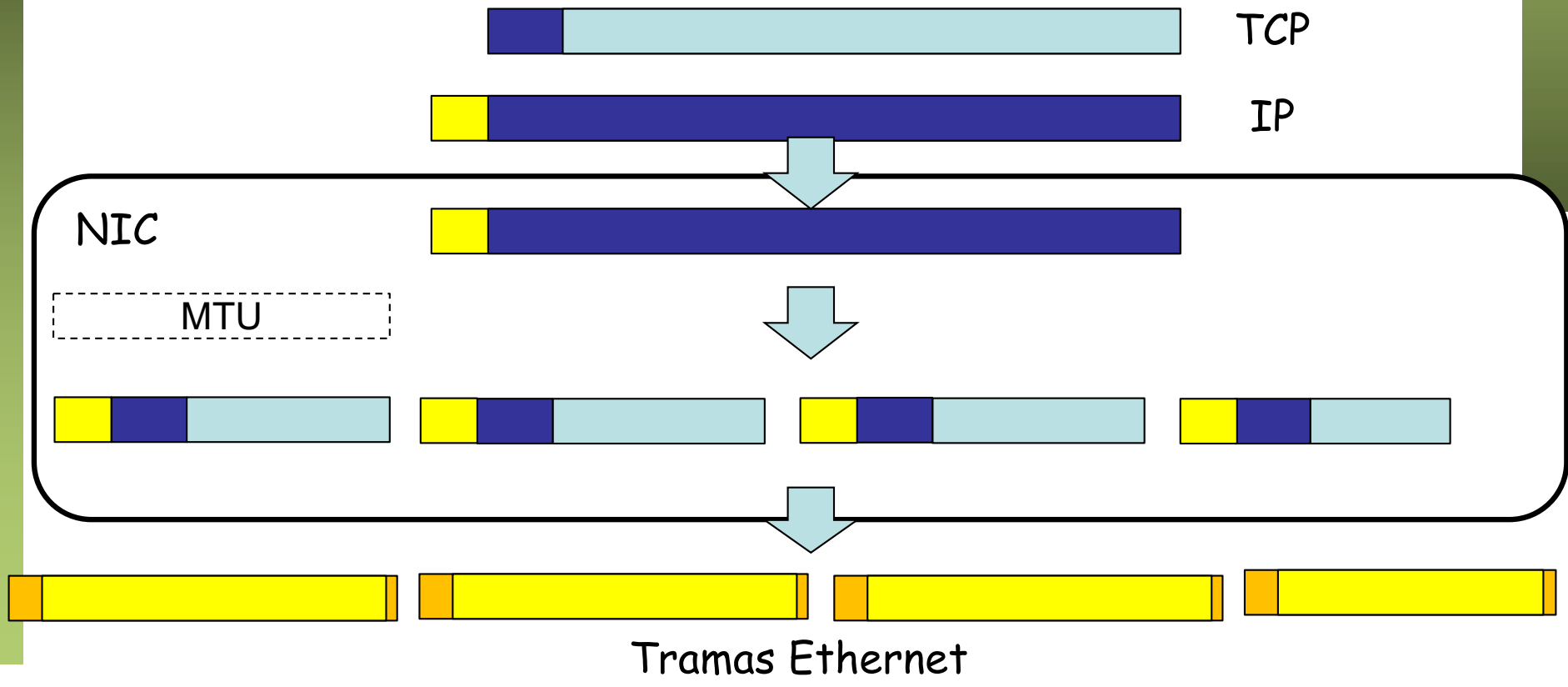
LSO

- *Large Segment Offload, TCP Segmentation Offload*
- Busca reducir carga de trabajo a la CPU en transmisión
- TCP entrega a la NIC paquetes más grandes que la MTU
- (...)



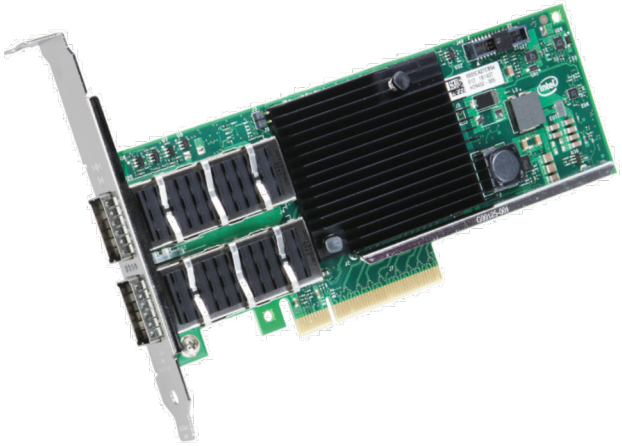
LSO

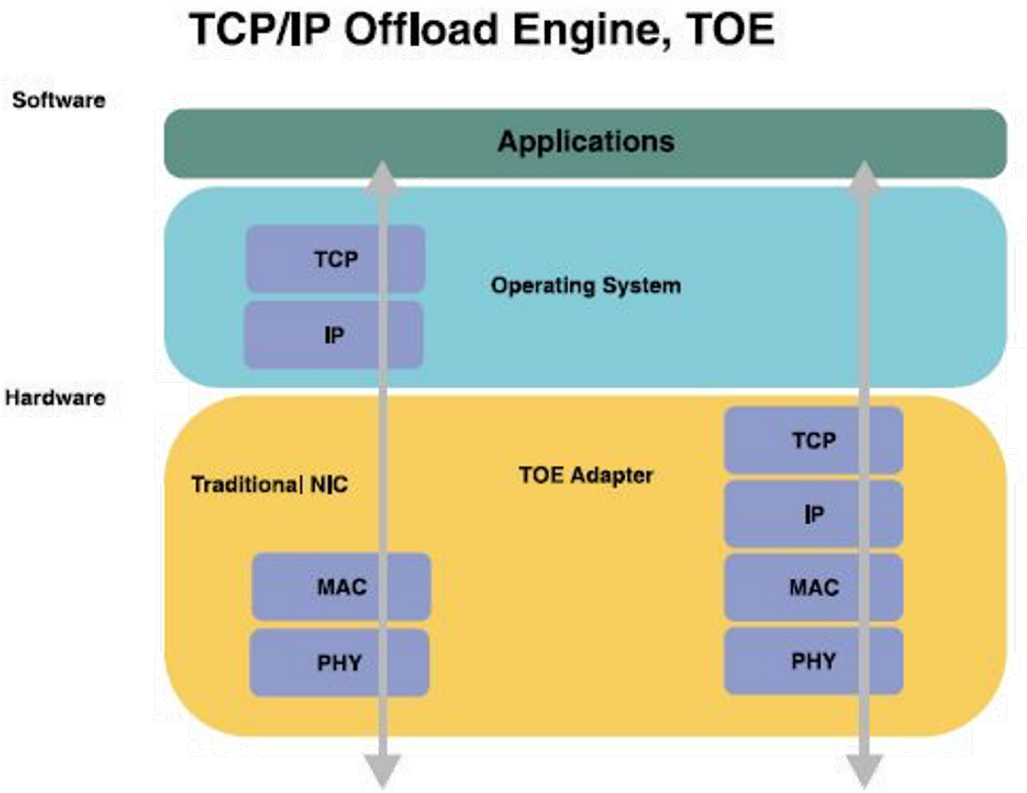
- La propia NIC hace la segmentación de nivel TCP
- Eso le obliga a crear nuevas cabeceras TCP e IP, descargando de ello a la CPU
- Requiere que la NIC sepa segmentar el protocolo (solo TCP)
- Problemas con encriptación (IPSec)
- Genera ráfagas de tráfico



TOE

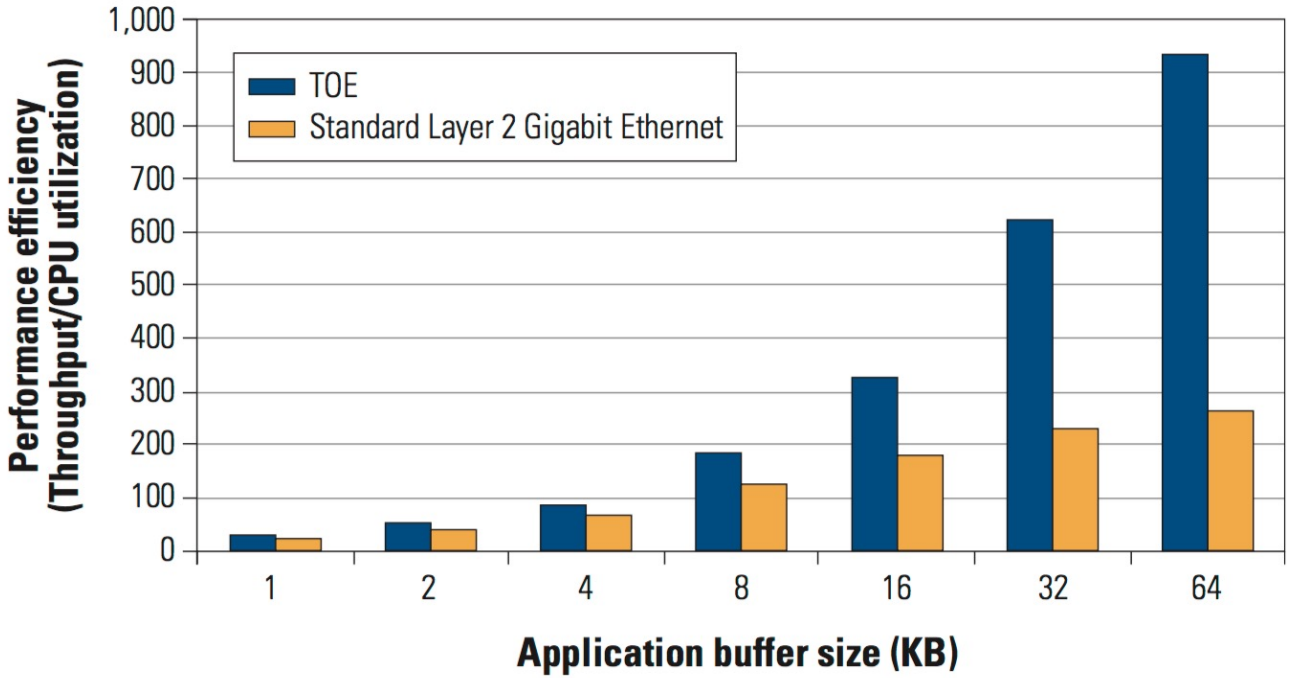
- *TCP/IP Offload Engine*
- Los datos pueden pasar directamente de la aplicación a la NIC
- La NIC puede emplearse para todas las tareas de la fase de transferencia y emplear la CPU para el establecimiento y terminación
- O se puede emplear la NIC para todo
- Requiere soporte del sistema operativo





TOE

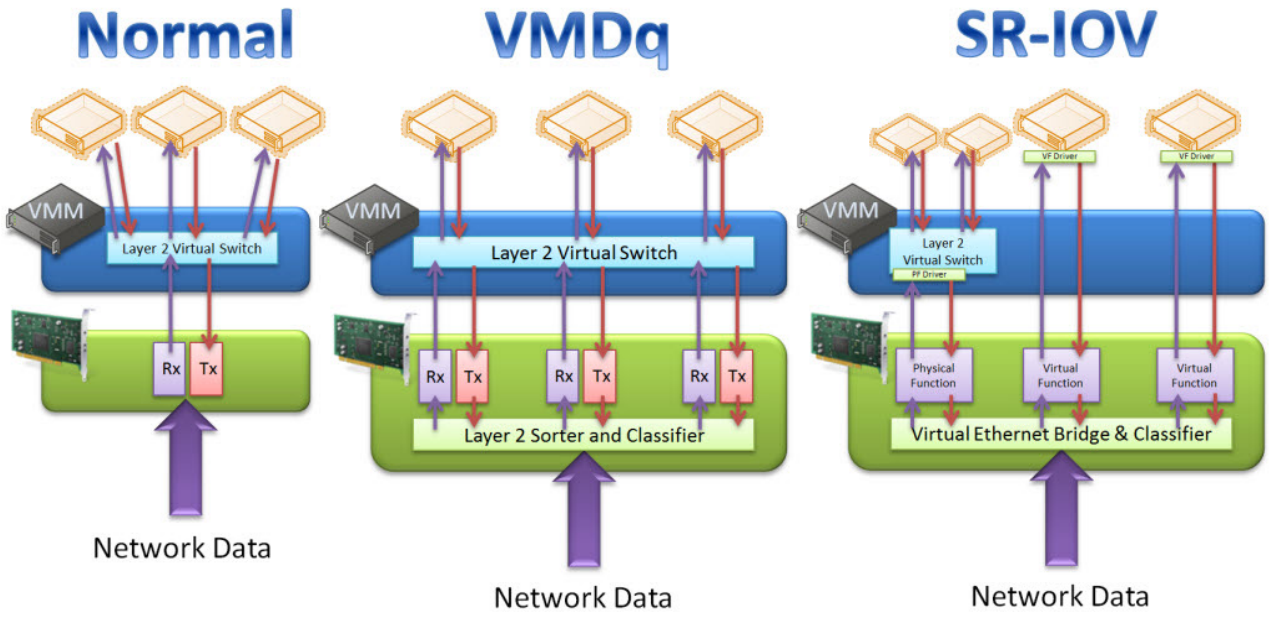
- Puede mejorar el throughput
- Reduce la carga sobre la CPU



http://www.dell.com/downloads/global/power/ps3q06-20060132-Broadcom.pdf

Otras funcionalidades

- VMDq, SR-IOV, etc, asociadas a la presencia de máquinas virtuales



http://www.consultia.co/what-is-vmdq/